فصل 2

​​​​​​​​

علم داده، آمار و سری زمانی

​​​​​​​​

پاتریک بنگرت

ایالات‌متحده، خیابان سامسونگ SDSA، شهر سن خوزه، تیم هوش مصنوعی؛ آلمان، خیابان بد نوهیم، خیابان کوچلر استراسه، پلاک ۷، تکنولوژی‌های الگوریتمی GmbH

در این فصل، مفاهیم، ابزار و گام‌هایی که برای تولید یک مجموعه داده تمیز و قابل ارائه که برای تجزیه تحلیل‌های بعدی آماده است را معرفی می‌کنیم. داشتن داده اینقدر ها هم ساده نیست که فقط با ثبت و ضبط داده‌های دوازده ماه گذشته سیستم‌های کنترلی بتوان داده‌های تمیز و خوبی بدست اورد و به تبع آن نتیجه خوبی هم نمی‌گیریم. در اینجا درباره تصمیماتی که قبل از استخراج داده باید گرفته شود و ترمیم و اصلاحاتی که روی داده بعد استخراج داده باید انجام شود به بحث و گفت و گو می پردازیم. بعد از انجام تمام این کارها در فصل بعد درباره تحلیل‌هایی که در مبحث یادگیری ماشین باید انجام دهیم صحبت می‌کنیم. هر چند دنبال کردن گام‌های این فصل برای آماده‌سازی داده‌ها بسیار زمان‌بر است و انرژی زیادی صرف می‌کند اما به شدت به احتمال موفقیت ما می افزاید.

داده‌هایی که به احتمال زیاد در صنعت نفت و گاز به آن بر می خوریم از جنس داده‌های عددی بوده و بر حسب زمان مرتب شده‌اند. به طور مثال دماسنجی که برای سال‌ها هر ده دقیقه دما را اندازه‌گیری کرده است. منبع این داده‌ها حسگری است که در میدان جای‌گذاری شده است می‌باشد. هرچند طی سالیان هر چند مدت یکبار این سنسور نیاز به تنظیم مجدد یا کالیبره شدن، تعمیر یا تعویض دارند اما زنجیره‌ی اندازه‌گیری هرگز متوقف نمیشود. جمع‌آوری این اندازگیری‌ها را سری زمانی می نامند. در پایگاه داده برچسبی که سری‌های زمانی را از دیگران تشخیص می‌دهد را معمولا برچسب می‌نامند. در ارائه و تحلیل سری‌های زمانی قدیمی‌تر نیازی به استفاده از یادگیری ماشین نبود (Hamilton 1994).

در هر واحد صنعتی، یا میدان نفتی، احتمالا به ده‌ها هزار برچسب بر بخوریم که دوره زمانی هر کدام از این‌ها می‌توانند تاریخچه‌ای در حد چندین سال داشته باشد. تناوب ضبط داده البته بسته به نوع و کاربرد آن می‌تواند از یکبار اندازه‌گیری در روز تا چندین اندازه‌گیری در ثانیه باشد.

عمده علم داده و یادگیری به انحاء مختلف در منابع و نوشته‌ها با داده گره خورده است. عکس‌ها، رشته کلمات، صوت ضبط شده و ویدیوها نمونه‌هایی از دیگر انواع داده‌ها هستند که ما کمتر در صنعت به آن‌ها برمی‌خوریم. این‌ها داده‌هایی هستند که در دسته وسیعی از نرم‌افزارها و اپلیکیشن‌ها که در خود از علم داده بهره برده‌اند استفاده می‌شوند و ما در این جا درباره آن‌ها صحبت نمی کنیم.

جدا از سری‌های زمانی، در صنعت به نوع دیگری داده نیز برمی‌خوریم به نام طیف. لرزش معمولا با معیاری از میزان حرکت در هر فرکانس اندازه‌گیری می‌شود. مجموعه این اندازه‌گیری‌ها را طیف لرزش می‌نامند. سپس آن‌ها را در بازه‌های زمانی منظم ثبت می‌کنند که خود یک سری زمانی را می‌سازد. سری‌های زمانی عادی از جنس عدد هستند اما این سری‌های زمانی همان طور که هر مورد داده‌های مشاهده شده و ثبت شده رشته ای از اعداد هستند از جنس بردار هستند. تحلیل‌های علمی مهمی برای طیف‌ها وجود دارد، از جمله تبدیل فوریه، که می توان از آن‌ها برای تجزیه تحلیل داده‌ها استفاده کرد (Press, Teukolsky, Vetterling, & Flannery, 2007).

عکس

حتی همین طور که به داده‌های خود نگاه کنیم، خود به خود به تدریج درک بالایی از آنچه این داده‌ها به ما می‌گویند پیدا می‌کنیم؛ اولیه‌ترین و سطحی‌ترین نوع تحلیل داده همین است. شکل 2.1 پنج مرحله اصلی سفر ماجراجویانه تحلیل داده را نشان می‌دهد. همانطور که می‌بینید در ابتدا، تنها چیزی که ما داریم مجموعه بزرگی از اعداد است، در واقع خود داده‌ها. حال در مرحله دوم، زمانی که همه داده‌های تکراری، پرت، اندازه گیری‌های به‌دردنخور و موارد مشابه حذف شوند و ما با داده های مهم و مرتبط باقی بمانیم، می‌توانیم بگوییم ما داده ها را به اطلاعات تبدیل کرده‌ایم. سوم، دانش زمانی به دست می‌آید که مشخص شود داده‌ها چگونه به هم متصل می‌شوند، چه برچسب‌هایی با برچسب‌های دیگر مرتبط هستند، این که چگونه این ارتباطات بر آنچه که ما در واقعیت تجربه می‌کنیم اثر می‌گذارد و غیره. چهارم، بینش زمانی ایجاد می‌شود که از طریق این تحلیل‌ها متوجه بشویم در چه نقطه‌ای هستیم و کجا باید باشیم. بعد از بینش گام پنجم رسیدن به خرد است؛ به این معنی که بدانیم چگونه به نقطه‌ای که باید باشیم برسیم. تنها چیزی که در این مرحله باقی می ماند اجرای طرح است.

این فصل برخی از تکنیک‌های رایج برخورد با داده‌های عددی در آماده‌سازی آن برای یادگیری ماشین را معرفی می‌کند (Pyle، 1999).

عکس

1.2 اندازه‌گیری، عدم قطعیت و نگهداری از سوابق

حسگرها کمیت‌ها را بر اساس نوعی پاسخ فیزیکی اندازه‌گیری می‌کنند و معمولاً آن را به سیگنال الکتریکی تبدیل می‌کنند. سپس با استفاده از مبدل آنالوگ به دیجیتال به عدد تبدیل می‌شود و به سیستم کنترل فرآیند ارسال می‌شود. این چرخه ممکن است بارها در یک ثانیه انجام شود. سیستم کنترل فرآیند عموما این داده ها را فقط برای مدت کوتاهی ذخیره می‌کند. وظیفه تاریخ‌نگار همین است که این که داده ها را برای درازمدت ذخیره کند و آنها را برای تجزیه و تحلیل، غالبا به صورت نموداری، آماده کند.

همانطور که در شکل 2.2 نشان داده شده است اندازه گیری برخی از کمیت‌ها، مثل دما، آسان است. اما اندازه‌گیری برخی از آنها دشوار است، مثل گرمای از دست رفته در محیط. نهایتا، اندازه‌گیری برخی از کمیت‌ها به صورت مستقیم غیرممکن است و باید بر اساس اندازه‌گیری برخی کمیت‌های دیگر، مانند آنتالپی، محاسبه شوند. برای اندازه گیری هر چیزی، باید حسگرهای مناسب آن را انتخاب کنیم و همچنین باید به درستی آن را در محل قرار دهیم تا خروجی های محسوس و منطقی به ما بدهد. فرض کنید دما را در زمان t=0، 1.5 درجه سانتیگراد، در زمان t=1، 3 درجه سانتیگراد و در زمان t=2، 2.5 درجه سانتیگراد اندازه‌گیری کرده‌ایم. با داشتن این داده‌ها در دست، دو سؤال اساسی به وجود می‌آید: 1) اندازه‌گیری که انجام شده است چقدر دقیق است، و (2) چگونه اندازه‌گیری‌ها را در تاریخ‌نگار ثبت کنیم؟

1.1.2 عدم قطعیت

هر حسگر درصد خطایی دارد؛ تبدیل آنالوگ به دیجیتال هم همینطور است. در طول مسیر از سنسور به مبدل مقداری از امواج از دست می‌رود و برخی از سیگنال‌های الکتریکی به‌درد نخور وارد می‌شوند. حسگرها در طول زمان دقت خود را از دست می‌دهند دقیقا مثل عقربه ترازو که صفر را نشان نمی‌دهد و باید در فواصل زمانی معین مجددا کالیبره یا تنظیم شوند. حسگرها گاها خراب می‌شوند و باید تعمیر شوند. تجمع مواد ممکن است حسگر را بپوشاند و اندازه گیری آن را از مقدار واقعی آن منحرف کند. دقت ذخیره اعداد در سیستم‌های فناوری اطلاعات، مثل سیستم های کنترل، با دقت محدود است و با هر تبدیل یا محاسبه، این دقت کم و کمتر می‌شود. فرقی نمی‌کند اندازه‌گیری چه چیزی، برای هر اندازه‌گیری این عوامل منجر به عدم قطعیت می‌شوند. ارزیابی عدم قطعیت کلی در هر اندازه‌گیری می‌تواند کاری بسیار دشوار باشد. در همان مثال دما، ممکن است تصمیم بگیریم که عدم قطعیت را ±1 درجه سانتیگراد در نظر بگیریم. این مسائل باعث می‌شوند که به طور مثال دو مشاهده آخر 3 و 2.5 اگر چه از نظر عددی متمایز هستند اما در واقع برابر باشند.

شکل

عدم قطعیت در محاسبات در اصطلاح علمی بدین معناست که محاسبات صورت گرفته بر اساس داده‌هایی انجام شده است که از دقت اندازه‌گیری آن‌ها مطمئن نیستیم. یافتن خود این عدم قطعیت که از محاسبات (بسته به خود محاسبه) یا اندازه‌گیری‌های قبلی به ارث رسیده است ممکن است دشوار باشد و می‌تواند خود منجر به عدم قطعیت بزرگتر از آنچه متخصصان و طراحان پیش بینی کرده‌اند، شود. به طور کلی، عدم قطعیت در یک کمیت مثل y که با استفاده از چندین ورودی نامطمئن مثل xi محاسبه شده است، با دیفرانسیل کل برابر است، اما در عمل استفاده از چنین معادله‌هایی برای محاسبه عدم قطعیت مشکل است بنابراین همان طور که در شکل 2.3 نشان داده شده است در بسیاری از موارد آن را تخمین می‌زنند.

|  |  |
| --- | --- |
| معادله |  |

هر کسی که تا به حال در آزمایشگاه کار کرده است و زمانی را در آنجا گذرانده است می‌ داند که تصحیح خطاهای اندازه‌گیری خاطره بسیار دردناکی است که ساعت‌ها عمر انسان را می‌گیرد و اگر قرار بود مسابقه‌ای میان همه کارهای زمان‌گیر پروژه مثل تحلیل داده‌ها و سایر آن‌ها برگذار شود، تصحیح خطای اندازه‌گیری رقیب پرمدعای قهرمانی بود. هنگامی که درگیر علم داده و تحلیل‌های مربوط به آن می‌شویم، نباید فراموش کنیم که این داده‌هایی که مبنای کار ما شده‌اند دقیق نیستند و بنابراین هر نتیجه‌گیری بر اساس آنها نیز می‌تواند لزوما دقیق نباشد. وقتی عدم قطعیت داده‌ها را محاسبه می‌کنیم و داده‌های خود راتصحیح می‌کنیم برخی اثرات و تحلیل‌ها که از قبل انجام داده بودیم ناگهان ناپدید می‌شوند!

به طور کلی، در غیاب اعمال اثرات عدم قطعیت‌ها، توصیه می‌کنیم هرگز به نتیجه گیری علم داده و تحلیل‌های آن اعتقاد نداشته باشید و بدان تکیه نکنید.

2.1.2 نگهداری و ضبط داده‌ها

تاریخ‌نگار اندازه‌گیری‌ها را در پایگاه داده ذخیره می‌کند. سیستم های IT یا فن‌آوری اطلاعات در طول زمان طراحی و در محل استقرار یافته‌اند تا در زمانی که ذخیره‌سازی در فضای دیسک‌های سخت گران و محدود بودند به کار گرفته شوند. تمایل داریم فضایی که داده‌ها اشغال می‌کنند را به حداقل برسانیم. برای این منظور یک راه و ایده خوب این است که خیلی ساده تعداد داده‌ها را کم کنیم و فقط مقادیری را ثبت کنیم که با آخرین مقدار ثبت شده بیش از حد معینی تفاوت داشته باشد. این مقدار به عنوان ضریب فشردگی شناخته می شود و یکی از ویژگی های ضروری هر برچسب است. معمولاً در زمان راه اندازی تاریخ‌نگار، در مورد ضریب فشردگی هر برچسب شخصی تصمیم می‌گیرد. اغلب یک مقدار پیش‌فرض برای آن مشخص می‌کنند و یا عمداً مقدار نسبتاً بزرگی به آن می‌دهند تا در فضای ذخیره‌سازی صرفه‌جویی کنند. شکل 2.4 وضعیت اندازه‌گیری دما که در بالا اشاره شد با ضریب فشردگی برابر با 1 را نشان می‌دهد.

شکل

اطلاعاتی که در نقاط جمع‌آوری داده به هر دلیلی ثبت نشوند برای همیشه از دست می‌روند. اگر ضریب فشردگی کمتر از عدم قطعیت باشد، در این صورت هیچ اطلاعات مفیدی در داده‌های از دست رفته وجود ندارد زیرا تمام آنچه که ثبت می‌کنیم نوسانات تصادفی خواهند بود. با این حال، اگر ضریب فشردگی بزرگتر از عدم قطعیت باشد، اطلاعات واقعی و مفید در مورد سیستم هم از دست می‌رود.

این یک راه حل ساده برای حل مسائل شرکت‌هایی که در تاریخ‌نگاری داده‌های خود به مشکل خورده‌اند بررسی و نگاهی عمیق‌تر به ضریب فشردگی است. در بسیاری از موارد، فرد متوجه می‌شود که ضریب بیش از حد بزرگ است و باید کمتر شود. این تغییر به تنهایی داده های بسیار ارزشمندتری را که از دست رفته بود در اختیار اپراتور قرار می‌دهد.

از آنجایی که ضریب فشردگی مانع از ثبت برخی اندازه‌گیری‌ها می‌شود. حال اگر تاریخ‌نگار قرار باشد مقدار اندازه‌گیری شده را برای زمانی که هیچ سابقه‌ای برای آن نداریم و اندازه‌گیری برای آن انجام نشده است، به عنوان خروجی به ما بدهد، آنگاه چه می‌کنیم؟ برای داشتن یک تصویر ذهنی خوب به شکل 2.5 مراجعه کنید. داده‌ای برای t = 3 وجود ندارد. چندین راه مختلف برای یافتن این مقدار و پاسخ به این سوال وجود دارد:

1. **پله‌کان**: مقدار برابر با آخرین عدد ثبت شده ثابت می‌ماند تا زمانی که اندازه‌گیری جدیدی صورت پذیرد.

2. **درون یابی**: از برخی توابع غیرخطی برای درون یابی داده‌ها استفاده می‌شود. معمولاً منحنی اسپلاین است.

3. **خطی**: یک خط مستقیم بین دو مشاهده متوالی رسم می‌شود تا داده‌ها را درون‌یابی کند. در واقع همان درون‌یابی خطی است.

شکل

هر چند که استدلال‌هایی هم موافق و هم بر ضد هر یک از این موارد بالا وجود دارد، اما توافق صنعت بر استفاده از روش اول، یا همان پله‌کان، به عنوان یک استاندارد است. با این دیدگاه یک برچسب با آخرین مقدار معتبر ثبت شده برای آن برابر است. مزیت قابل ملاحظه این روش این است که داده‌های ما در یک قالب برخط نگه داشته می‌شود، بدین معنا که در مواقعی که در هر لحظه نیاز به دانستن یک مقدار داشته باشیم بدون دانستن مقدار بعدی که در سری زمانی قرار است اندازه‌گیری شود می‌توانیم آن را بدست آوریم.

مبنای تحلیل مجموعه داده‌های موجود، شکل دهی صحیح به عملکرد و پیکربندی درست تاریخ‌نگار داده‌ها است. اگر یک مجموعه داده آماده به ما ارائه شود، بدون داشتن اطلاعاتی درباره عدم قطعیت داده‌ها، این گونه فرض می‌کنیم که عدم قطعیت هر تگ معمولاً به طور کامل با ضریب فشردگی پوشیده می‌شود. بنابراین در عمل، ما متخصصان علم داده احتمالا خواهیم گفت که عدم قطعیت همان ظریب فشردگی است و با هم برابر هستند. با این وجود، ضریب فشردگی باید توسط مهندسان فرآیند مورد بررسی قرار گیرد.

2.2 همبستگی و مقیاس زمانی

اگر یک برچسب مقدار خود را تغییر دهد، ما می‌توانیم چیزهایی در مورد میزان تغییر در مقدار برچسب‌های دیگر بگوییم. مثلا، با افزایش دما، انتظار داریم که فشار نیز افزایش یابد. این دو برچسب را مرتبط یا همبسته می‌نامیم، به شکل 2.6 مراجعه کنید. تغییر در برچسب اول ممکن است اطلاعاتی در مورد تغییرات وارده در برچسب دوم را نشان دهد، اما نه اینکه گویای همه چیز باشد. معیاری که میزان اطلاعات ارائه شده را نشان می دهد ضریب همبستگی است که عددی بین 1- و 1 است. اعداد مثبت نشان می‌دهند که با افزایش برچسب اول، برچسب دوم نیز افزایش می‌یابد. اعداد منفی نشان می‌دهند که با افزایش برچسب اول، برچسب دوم کاهش می یابد. عدد 1، چه مثبت چه منفی نشان می دهد که تغییر در هر دو برچسب با یک گام و میزان یکسان اتفاق می افتد. هر چقدر تغییر مرتبط باشد، این دو برچسب بیشتر به هم نزدیک و یکسان هستند. مقدار 0 نشان می‌دهد که این دو برچسب اصلاً مرتبط نیستند. تغییر در برچسب اول هیچ اطلاعاتی در مورد تغییر در برچسب دوم ارائه نمی‌کند.

در صنعت نفت و گاز مجموعه داده‌هایی که به دست ما می‌رسد از بسیاری از بخش های به هم پیوسته حاصل می‌شوند. بدین معنا که در این صنعت بسیاری از برچسب‌ها، مانند دما و فشار یک مکان، همبسته هستند. تجزیه و تحلیل اینکه کدام مقادیر بیشتر همبسته هستند و یا کمتر، به طور طبیعی یک حالت خوشه بندی جالب در برچسب‌ها به وجود می‌آورد.

به طور کلی، درک ما به ما می‌گوید همبستگی در قالب وابستگی خطی اندازه‌گیری می‌شود که در شکل 2.6 نیز نشان داده شده است. معیاری که معمولاً برای نشان دادن قدرت یک یک رابطه خطی استفاده می‌شود، ضریب همبستگی پیرسون است. اگر رابطه ما غیر خطی باشد، ممکن است استفاده از این ضریب همبستگی برای آن به هیچ وجه معنی‌دار نباشد. از آنجایی که بسیاری از فرآیندهای فیزیکی غیر خطی هستند در نظر گرفتن این قضیه خیلی مهم است. تحلیل همبستگی غیرخطی پیش‌فرض می‌گیرد که نوع غیرخطی موجود را بشناسیم، و این یک چالش دیگر است. در عمل، ضریب همبستگی خطی را منحصراً مورد استفاده قرار می‌دهیم. این خوب است اگر به درستی درک شود.

از آنجایی که ما با سری‌های زمانی سروکار داریم، مهم است که یادآور شویم که همبستگی را معمولاً برای مقادیری محاسبه می کنیم که در زمان یکسان ثبت و ضبط شده باشند. اگر یکی از دو سری زمانی را چند گام از لحاظ زمانی جابجا کنیم، تابع همبستگی، که همان ضریب همبستگی به عنوان تابعی از زمان جا به جا شده است را بدست می آوریم. یک حالت خاص آن این است که همبستگی یک برچسب را به خودش در حالتی روش جا به جایی زمان را روی آن اعمل کرده‌ایم محاسبه کنیم که به این حالت خاص تابع خودهمبسته می‌گویند. این توابع احتمالا شکلی شبیه به شکل 2.7 دارند.

همانطور که مشاهده می‌کنیم برای تاخیر زمانی برابر صفر همبستگی بسیار بالاست، سپس کاهش می‌یابد تا در جا به جایی زمانی برابر 10 به یک مقدار حداقل می‌رسد و در 20 دوباره بالا می‌آید تا به حداکثر خود برسد و به همین شکل پیش می‌رود. خوب به وضوح می‌بینیم که این دو سری زمانی با مکانیزمی که فعلا نمی‌دانیم با 10 گام تاخیر زمانی با هم در اتباط هستند. آن زمانی که در ذات دینامیکی سیستم قابل مشاهده باشد را مقیاس زمانی سیستم می‌نامیم. بسیار بسیار مهم است که